Probability and Statistical Inference
Continuous Assessment Part II

Student Number: D14122773
Name: Yahia Ragab
Programme Code: TU060/DS

Section 1 - Research Question(s)

This paper aims to use a dataset to answer a research question as well as hypotheses stated prior to conducting the analysis on the data. This will be done by, first, exploring and understanding the dataset, preparing the dataset to be used in a Linear Regression Model. The exploration of the dataset will demonstrate statistics about the sample's numeric and categorical variables, as well as identify the variables of interest to select which variables would be best to use in the Linear Regression model. The preparation will include ensuring that the assumptions of the variables used in the Linear Regression are all met.
The research question that this paper will set out to answer is:

How far can we reduce the dimensionality of the dataset while retaining the variability in the target variable (Global Score)?

The paper aims to answer this question by investigation the relationship between a student's Global Score in the Saber PRO exam (G_SC), the student's academic performance in high school and in university, as well as the student's socio-economic and parental situation.

Hypotheses Prior to the Research

1. Not all 15 variables affect the target variable
2. Citizen Competencies will not have an effect on the the Global Score
3. The English Score in Saber Pro (University) is highly dependent on the English Score in Saber 11 (Grade 11)
4. A student's social class affects his/her ability to speak English positively

Section 2 – Dataset

The dataset which will be used in this project is related to the secondary and university education assesments for students in the engineering field and contains social, economic, academic data for all 12,411 students. The dataset contains 45 variables which are shown in the table below:

Table 1
Description of numeric variables.

Variable	Full name	Mean	Standard Deviation	Max	Min
MAT_S11	Mathematics	64.32	11.87	100	26
CR_S11	Critical Reading	60.78	10.03	100	24
CC_S11	Citizen Competencies S11	60.71	10.12	100	0
BIO_S11	Biology	63.95	11.16	100	11
ENG_S11	English	61.80	14.30	100	26
QR_PRO	Quantitative Reasoning	77.42	22.67	100	1
CR_PRO	Critical Reading	62.20	27.67	100	1
CC_PRO	Citizen Competencies SPRO	59.19	28.99	100	1
ENG_PRO	English	67.50	25.49	100	1
WC_PRO	Written Communication	53.70	30.00	100	0
FEP_PRO	Formulation of Engineering Projects	145.48	40.12	300	1
G_SC	Global Score	162.71	23.11	247	37
PERCENTILE	Percentile	68.45	25.87	100	1
2ND_DECILE	Second Decile	3.89	1.25	5	1
QUARTILE	Quartile	3.19	0.98	4	1
SEL	Socioeconomic Level	2.60	1.11	4	1
SEL_IHE	Socioeconomic Level of The Institution of Higher Education	2.41	0.93	4	1

Note: *S_11 corresponds to the secondary test and S_PRO to the professional test.*

Probability and Statistical Inference
Continuous Assessment Part II

Table 2
Description of categorical variables.

Variable	Full Name	Levels	Variable	Full name	Levels
GENDER	Gender	2	DVD	DVD	2
EDU_FATHER	Father's education	12	FRESH	Fresh	2
EDU_MOTHER	Mother's education	12	PHONE	Phone	2
OCC_FATHER	Father's occupation	13	MOBILE	Mobile	2
OCC_MOTHER	Mother's occupation	13	REVENUE	Revenue	3
STRATUM	Stratum	7	JOB	Job	8
SISBEN	Sisben	6	SCHOOL_NAME	School name	3.735
PEOPLE_HOUSE	People in the house	13	SCHOOL_NAT	Nature of School	2
INTERNET	Internet	2	SCHOOL_TYPE	Type of School	4
TV	TV	2	COD_SPRO	Code Saber Pro	12.411
COMPUTER	Computer	2	UNIVERSITY	University	134
WASHING_MCH	Washing machine	2	ACADEMIC_PROGRAM	Academic Program	23
MIC_OVEN	Microwave oven	2	COD_S11	Code Saber 11	12.411
CAR	Car	2			

Variables of Interest

The target variable (Y) is the Global Score (G_SC) as it is representative of the student's academic performance. The predictor variables to be used are all the student's academic and personal skills, which are the numeric variables available in the dataset: MAT_S11, CR_S11, BIO_S11, ENG_S11, QR_PRO, CR_PRO, CC_PRO, ENG_PRO, WC_PRO, FEP_PRO, SEL, SEL_IHE. These do not include PERCENTILE, QUARTILE and 2ND DECILE as these are all derived variables. Categorical variables were not used in this analysis.

Missing Variables

The first step taken in exploring the data was checking if the dataset contains any missing data. One variable named "..10" was found to have 12,411 missing values, which was not present in the data description. Upon inspection, this variable turned out to be the result of the CSV file downloaded containing an empty column, which was read in by R as variable. This variable was removed. No missing data was found in the dataset.

Data Exploration

This section will investigate graphical representations of the data to understand and identify any patterns between the data, as well as to identify if any transformation is required prior to using the variables in the Linear Regression model.

Categorical Variables

Variable	Number of Levels	Variable	Number of Levels
w	12411	**CAR**	2
GENDER	2	**DVD**	2
EDU_FATHER	12	**FRESH**	2
EDU_MOTHER	12	**PHONE**	2
OCC_FATHER	12	**MOBILE**	2
OCC_MOTHER	12	**REVENUE**	8
STRATUM	7	**JOB**	4
SISBEN	6	**SCHOOL_NAME**	3651
PEOPLE_HOUSE	13	**SCHOOL_NAT**	2
INTERNET	2	**SCHOOL_TYPE**	4
TV	2	**Cod_SPro**	12395
COMPUTER	2	**UNIVERSITY**	134
WASHING_MCH	2	**ACADEMIC_PROGRAM**	21
MIC_OVEN	2		

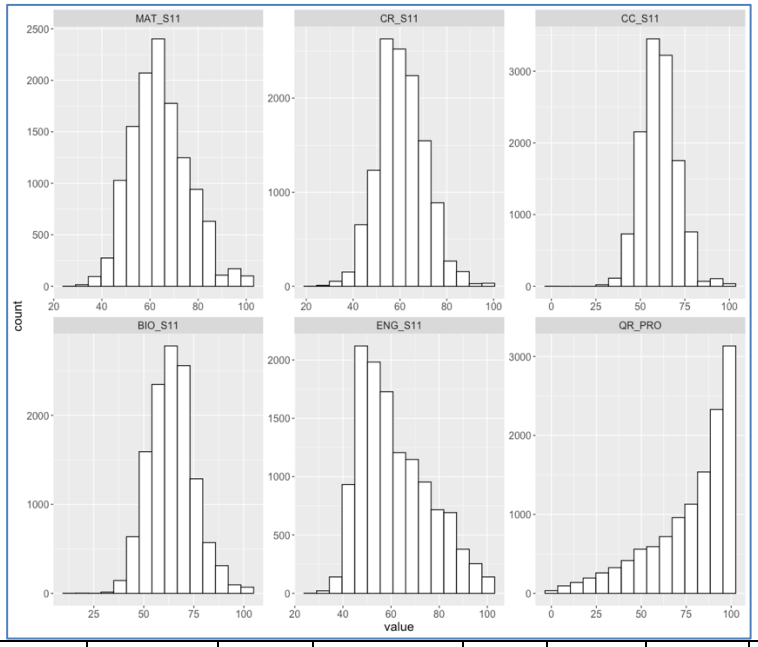
Probability and Statistical Inference
Continuous Assessment Part II

Variable	Freq	Variable	Freq
COD_S11	SB11201210000129: 1	COMPUTER	No : 2237
	SB11201210000137: 1		Yes:10174
	SB11201210005154: 1	WASHING_MCH	No :4723
	SB11201210007504: 1		Yes:7688
	SB11201210007548: 1	MIC_OVEN	No :3841
	SB11201210007568: 1		Yes:8570
	(Other):12405	CAR	No :6602
GENDER	F:5043		Yes:5809
	M:7368	DVD	No :3089
EDU_FATHER	Comp prof ed:3016		Yes:9322
	Complete Secundary: 2843	FRESH	No : 381
	Comp tech or tech:1194		Yes:12030
	Incomplete Secundary :1091	PHONE	No : 521
	Postgraduate education:1085		Yes:11890
	Complete primary: 824	MOBILE	No :3564
	(Other):2358		Yes:8847
	Complete Secundary: 3106	REVENUE	Between 1 and less than 2 LMMW:3873
	Comp prof ed:3059		Between 2 and less than 3 LMMW:2783
	Comp tech or tech: 1495		Between 3 and less than 5 LMMW:2239
	Incomplete Secundary: 1056		less than 1 LMMW: 1037
	Postgraduate education: 997		Between 5 and less than 7 LMMW: 973
	Complete primary: 713		10 or more LMMW: 718
	(Other):1985		(Other): 788
OCC_FATHER	Independent:2907	JOB	0.095833333
	Tech or prof level emp:1803		No: 11909
	Operator: 1537		Yes, 20 hours or more per week: 134
	Other occupation: 1087		Yes, less than 20 hours per week: 230
	Executive: 1077	SCHOOL_NAME	CIUDAD ESCOLAR DE COMFENALCO: 47
	0.652777778		COL LA SALLE: 42
	(Other): 3060		COLEGIO DEL SAGRADO CORAZON: 40
	Home : 4658		COL. MILITAR ALMIRANTE COLON: 39
	Tech or prof level em:1795		COL CALASANZ: 38
	Independent: 1107		COL CHAMPAGNAT: 33
	Auxiliary or Admini: 846		(Other): 12172
	Executive: 794	SCHOOLNAT	PRIVATE:6565
	Independent professional: 715		PUBLIC :5846
	(Other): 2496	SCHOOLTYPE	ACADEMIC: 7834
STRATUM	0:14		Not apply: 5
	Stratum 1:1709		TECHNICAL: 1059
	Stratum 2:4029		TECHNICAL/ACADEMIC: 3513
	Stratum 3:4045	Cod_SPro	EK201830012603: 2
	Stratum 4:1578		EK201830013197: 2
	Stratum 5: 633		EK201830017763: 2
	Stratum 6: 403		EK201830022057: 2
SISBEN	0:21		EK201830030238: 2
	Esta clasificada…: 96		EK201830036216: 2
	It is not classified.…: 7534		(Other) :12399
	Level 1: 2057	UNIVERSITY	UNIVERSIDAD DE LOS ANDES-…: 696
	Level 2: 2120		ESCUELA COLOMBIANA …: 397
	Level 3: 583		UNIVERSIDAD INDUSTRIAL …: 397
PEOPLE_HOUSE	Four: 4767		UNIVERSIDAD DEL NORTE BARRANQUILLA: 376
	Five: 2870		UNIVERSIDAD DISTRITAL"... : 335
	Three :2345		FUNDACION UNIVERSIDAD DE .. : 326
	Six: 1090		(Other): 9884
	Two: 592	ACADEMIC_PROGRAM	INDUSTRIAL ENGINEERING:5318
	Seven : 372		CIVIL ENGINEERING :3320
	(Other): 375		MECHANICAL ENGINEERING:1135
INTERNET	No: 2659		CHEMICAL ENGINEERING :1000
	Yes:9752		ELECTRONIC ENGINEERING: 849
TV	No : 1842		ELECTRIC ENGINEERING : 278
	Yes:10569		(Other): 511

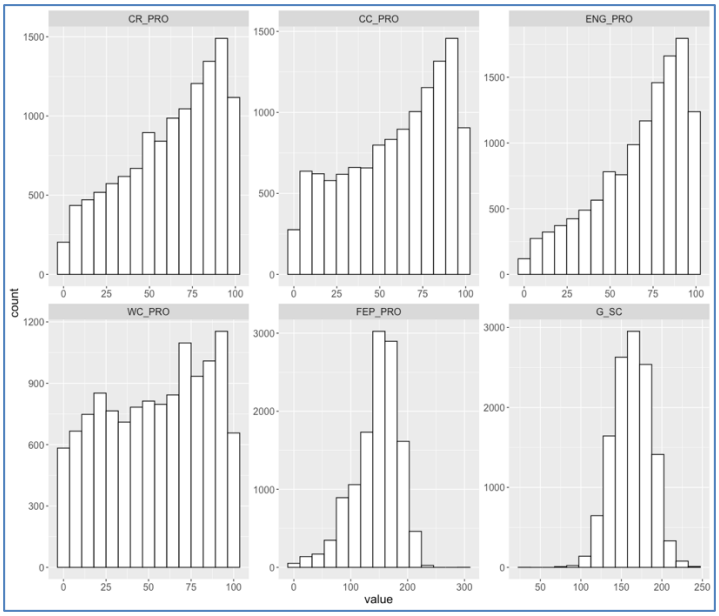
Probability and Statistical Inference
Continuous Assessment Part II

Central Tendency, Shape, and Variance: Histograms

The central tendency, shape and variance of numeric variables will be explored by visually inspecting histograms. Histograms are useful for showing patterns that the data follows. These are known as distributions. When a histogram has a bell shape, with the same mean, median and mode, it is called a normal distribution. It can also be skewed to one side, meaning that the mode is on the right or the left. The following section will show histograms and the minimum, 1st quartile, mean, 3rd quartile, maximum, range, variance and standard deviation. These are used to indicate the spread of the data.

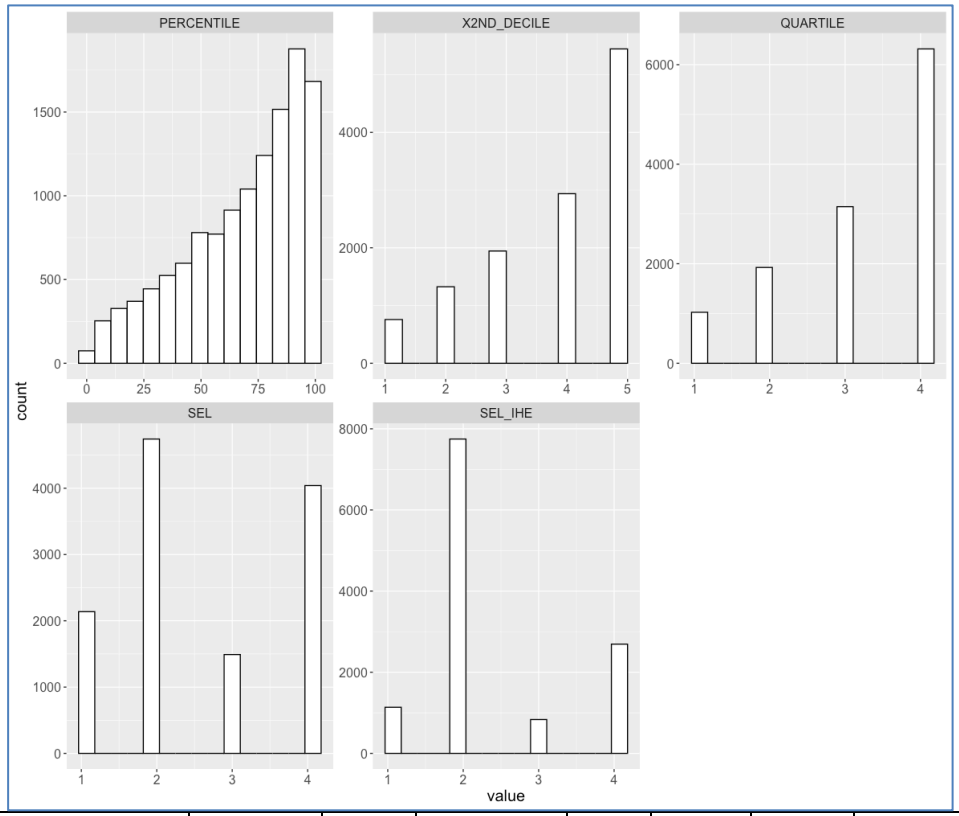


	min	25%	mean	75%	max	range	var	stddev
MAT_S11	26	56	64.32	72	100	74	140.98	11.87
CR_S11	24	54	60.78	67	100	76	100.52	10.03
CC_S11	0	54	60.71	67	100	100	102.43	10.12
BIO_S11	11	56	63.95	71	100	89	124.48	11.16
ENG_S11	26	50	61.8	72	100	74	204.43	14.3
QR_PRO	1	65	77.42	96	100	99	514.09	22.67



Probability and Statistical Inference
Continuous Assessment Part II

	min	25%	mean	75%	max	range	var	stddev
CR_PRO	1	42	62.2	86	100	99	765.44	27.67
CC_PRO	1	36	59.19	85	100	99	840.53	28.99
ENG_PRO	1	51	67.5	88	100	99	650	25.5
WC_PRO	0	28	53.7	80	100	100	900.1	30
FEP_PRO	1	124	145.48	174	300	299	1610.13	40.13
G_SC	37	147	162.71	179	247	210	534.19	23.11

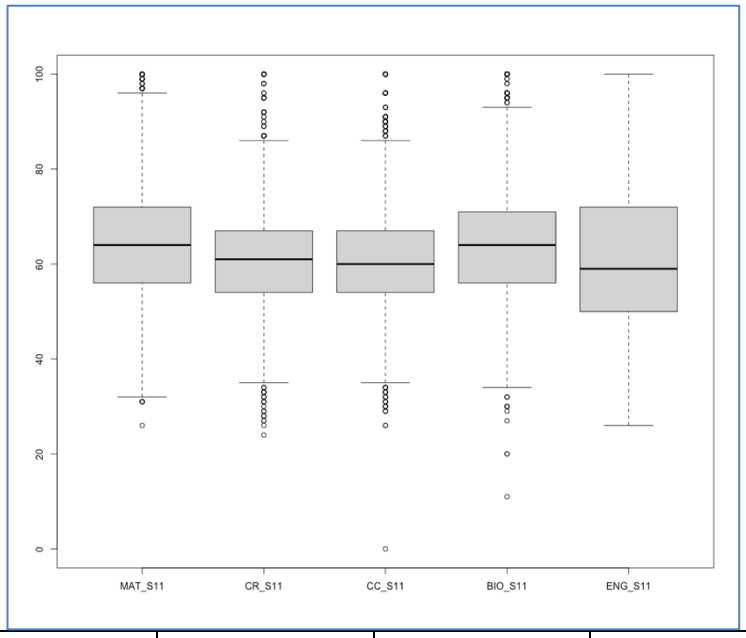


	min	25%	mean	75%	max	range	var	stddev
PERCENTILE	1	51	68.45	90	100	99	669.13	25.87
X2ND_DECILE	1	3	3.89	5	5	4	1.56	1.25
QUARTILE	1	3	3.19	4	4	3	0.96	0.98
SEL	1	2	2.6	4	4	3	1.24	1.11
SEL_IHE	1	2	2.41	3	4	3	0.86	0.93

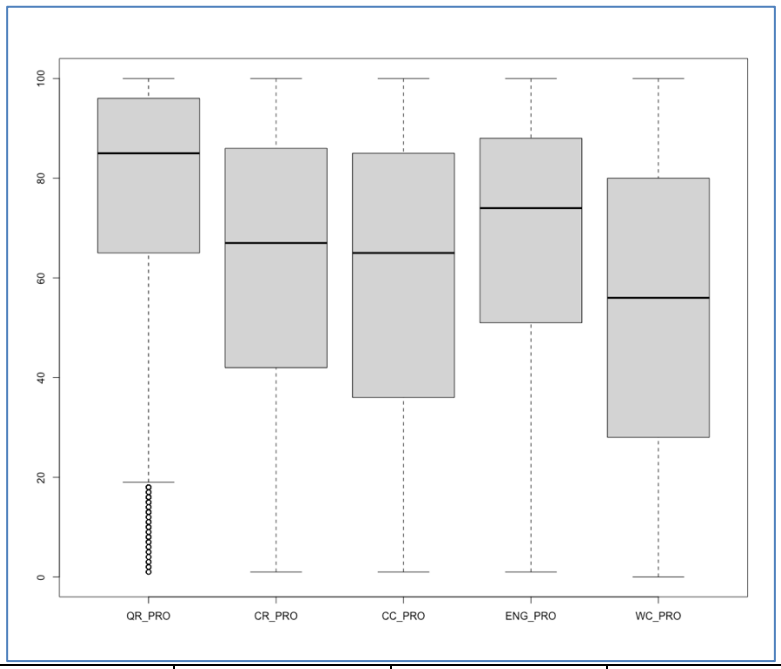
Outliers and Spread: Boxplots

When exploring numeric variables it is important to understand the spread of the data and to detect any outliers – unusual values in a variable that fall outside the range. Boxplots are very useful visualisations to display outliers. They display whiskers that indicate the normal range of the data, and the boxes show the 1st quartile, median, and 3rd quartile. The numeric values of these graphs are shown in the tables below the graphs.

Probability and Statistical Inference
Continuous Assessment Part II

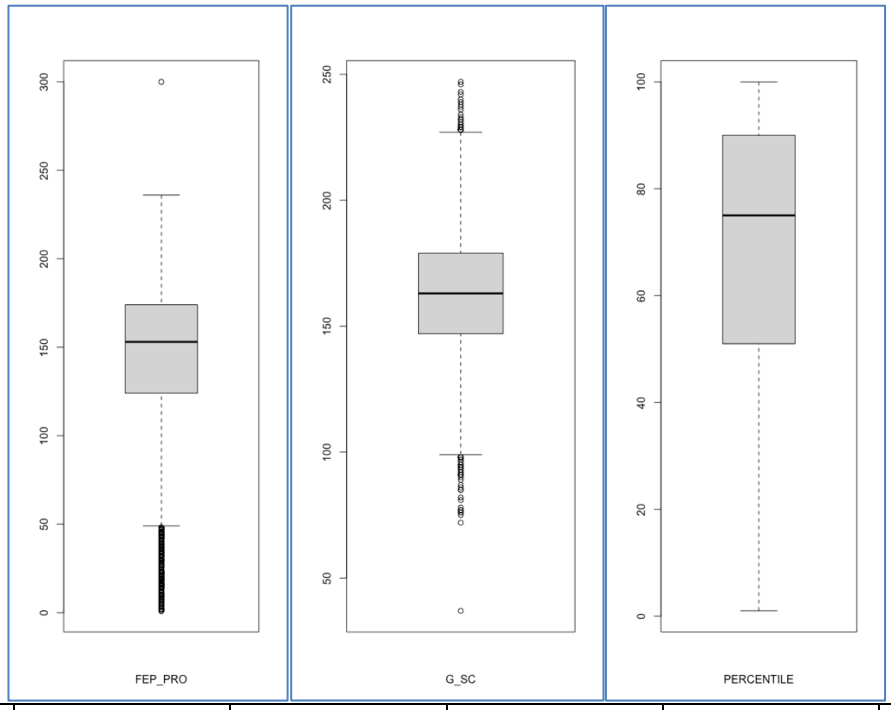


Variable	Lower Whisker	Lower Hinge	Median	Upper Hinge	Upper Whisker
MAT_S11	32	56	64	72	96
CR_S11	35	54	61	67	86
CC_S11	35	54	60	67	86
BIO_S11	34	56	64	71	93
ENG_S11	26	50	59	72	100

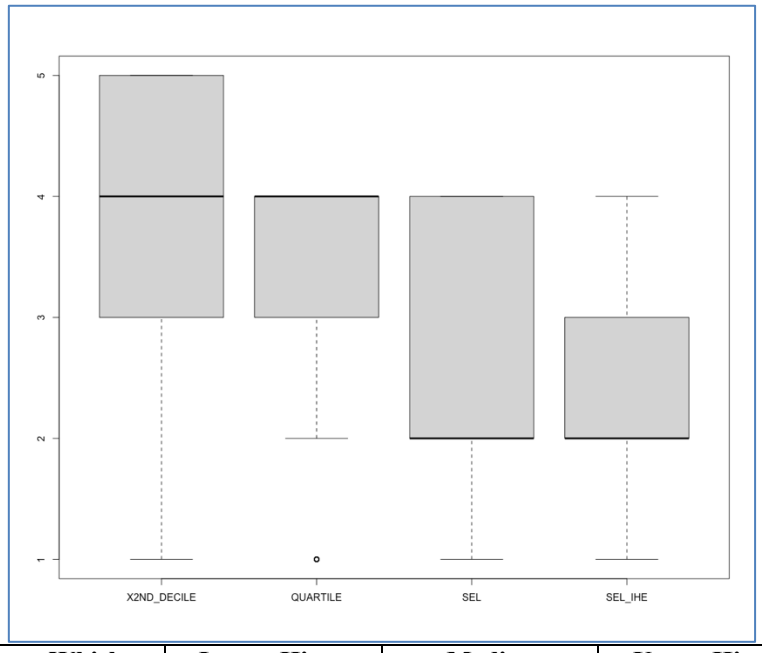


Variable	Lower Whisker	Lower Hinge	Median	Upper Hinge	Upper Whisker
QR_PRO	19	65	85	96	100
CR_PRO	1	42	67	86	100
CC_PRO	1	36	65	85	100
ENG_PRO	1	51	74	88	100
WC_PRO	0	28	56	80	100

Probability and Statistical Inference
Continuous Assessment Part II



Variable	Lower Whisker	Lower Hinge	Median	Upper Hinge	Upper Whisker
FEP_PRO	49	124	153	174	236
G_SC	99	147	163	179	227
PERCENTILE	1	51	75	90	100



Variable	Lower Whisker	Lower Hinge	Median	Upper Hinge	Upper Whisker
X2ND_DECILE	1	3	4	5	5
QUARTILE	2	3	4	4	4
SEL	1	2	2	4	4
SEL_IHE	1	2	2	3	4

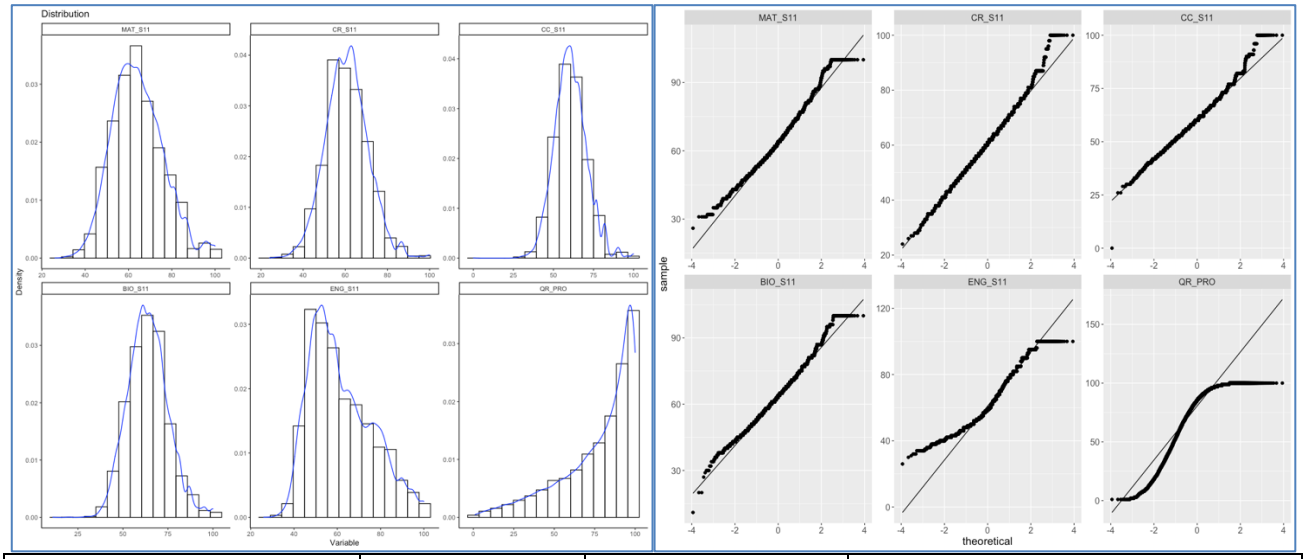
Most of the graphs showed many outliers which may affect the Linear Regression model. The standardised scores will be inspected for these values to see the percentage of values that falls outside of the range ±3.29.

Probability and Statistical Inference
Continuous Assessment Part II

Assessing Normality: QQ Plots and Density Curves

The numeric variables in the dataset were assessed for normality using multiple measures. First, the distribution of each variable was presented graphically in the form of a ***QQ Plot*** and a ***Density Curve***. Normally, the second step would be to use tests that assess normality such as the ***Shapiro-Wilk*** or ***Kolmogorov Smirnov***, and to check if their p-value. For a significance level of 5%, if the p-value < 0.05, the null hypothesis that the variable is normal is rejected, i.e. the variable does not follow a normal distribution. The limitation, however, with the Shapiro-Wilk and Kolmogorov Smirnov tests are that they do not work well with large dataset as they are too sensitive, and almost always show that the variable is normal in R.

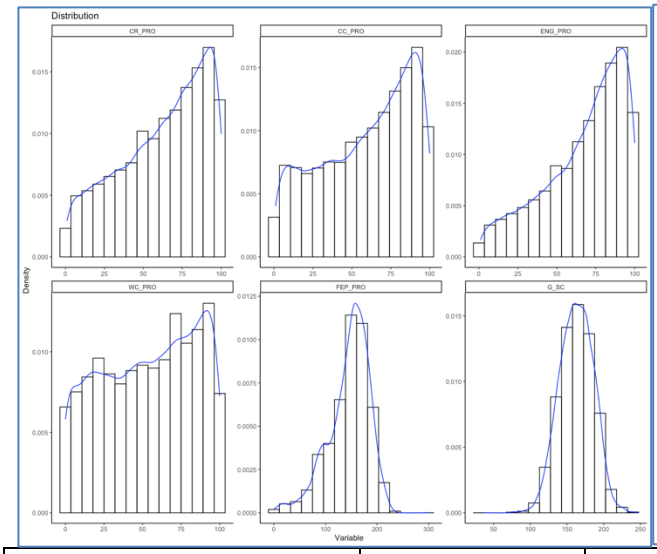
Instead, ***Skewness*** and ***Kurtosis*** were quantified and their standardised scores were checked. The skewness value for a normal distribution is 0 and for kurtosis it is 3, although it is often subtracted by 3 to make the normal point 0, this is called *zero kurtosis*. The acceptable range for skewness and zero kurtosis is ±2. Finally, the percentage of values in the variable that falls outside of the range ±3.29 is checked, and if it is less than 6, then the variable can be treated as normal. It is also an indication that the outliers are not affecting the variable. If the variable is found to be non-normal, non-parametric tests need to be used.

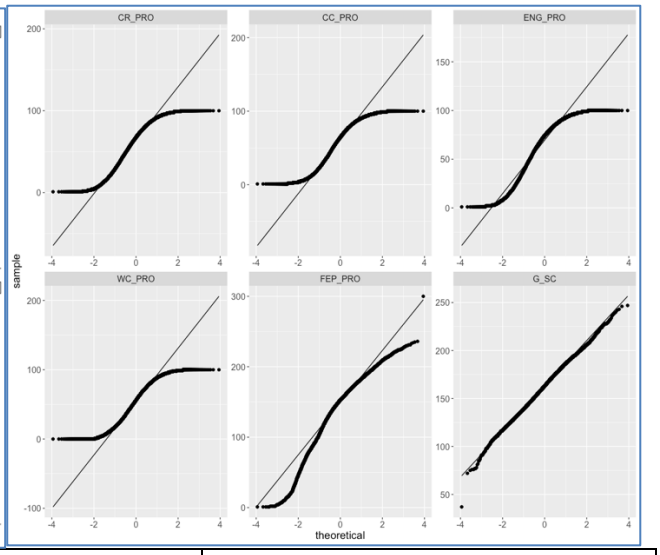


	kurtosis	skewness	PercOutside_3.29
MAT_S11	0.13	0.4	0
CR_S11	0.48	0.21	0.38
CC_S11	0.76	0.35	0.44
BIO_S11	0.3	0.3	0.03
ENG_S11	0.37	0.61	0
QR_PRO	0.63	-1.18	0.18

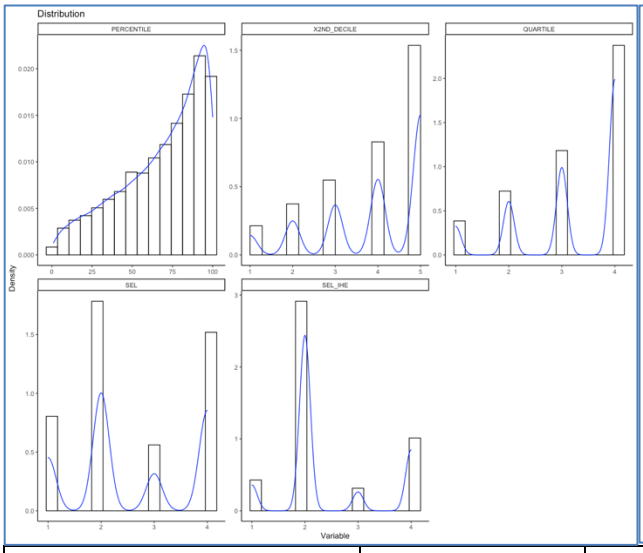
Upon inspecting the graphs, ENG_S11 and QR_PRO stand out as the two variables that are skewed and contain kurtosis. The standardised skewness and kurtosis scores for ENG_S11 showed that it does contain positive skewness, and kurtosis. The scores showed QR_PRO to be highly negatively skewed and to contain kurtosis. Both variables, however, had had lower than 6% of their values outside the range ±3.29, therefore the variables can be treated as normal.

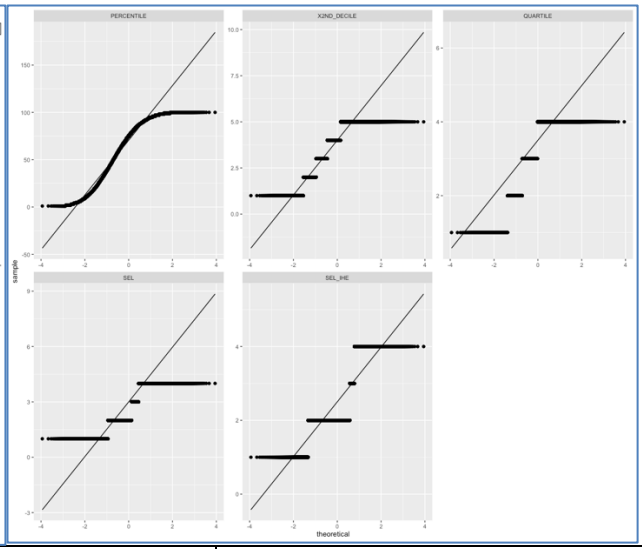
Probability and Statistical Inference
Continuous Assessment Part II





	kurtosis	skewness	PercOutside_3.29
CR_PRO	-0.85	-0.52	0
CC_PRO	-1.06	-0.42	0
ENG_PRO	-0.32	-0.8	0
WC_PRO	-1.2	-0.18	0
FEP_PRO	0.72	-0.84	0.51
G_SC	-0.08	-0.1	0.15





	kurtosis	skewness	PercOutside_3.29
PERCENTILE	-0.46	-0.75	0
X2ND_DECILE	-0.38	-0.87	0
QUARTILE	-0.36	-0.91	0
SEL	-1.41	0.09	0
SEL_IHE	-0.59	0.74	0

Percentile, 2nd Decile, Quartile, SEL, SEL_IHE score data was assessed for normality. Visual inspection of the histogram and QQ-Plot showed possible skewness and kurtosis. The standardised score for kurtosis (-0.46, -0.38, -0.36, -1.41, -0.59, respectively) and the standardised score for skewness (-0.75, -0.87, -0.91, 0.09, 0.74, respectively) were outside the acceptable range. However 100% of these values, standardised, fall within the bounds of +/- 3.29, thus the data can be considered to approximate a normal distribution.

Probability and Statistical Inference
Continuous Assessment Part II

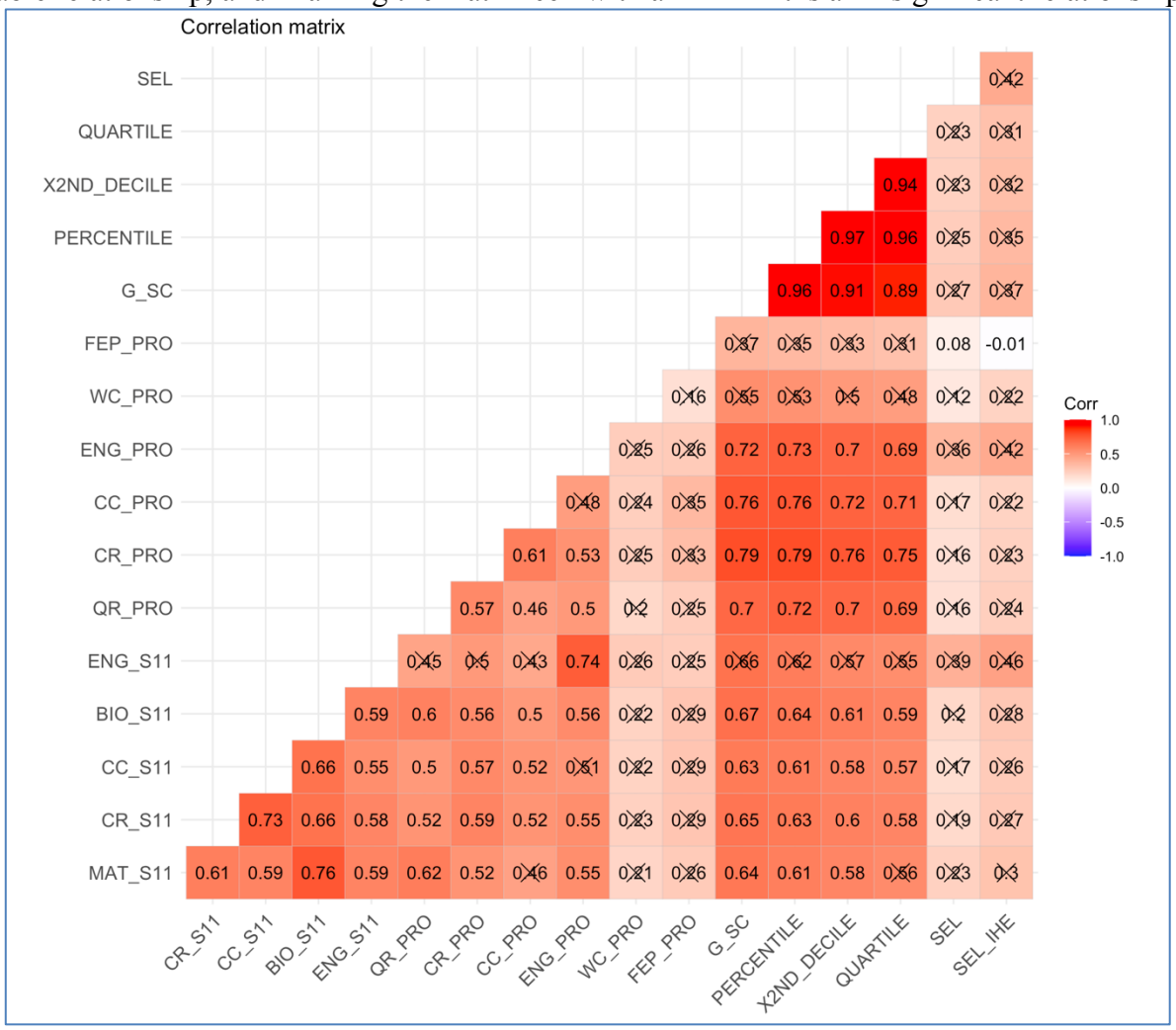
Option B

Section 3.1 Dimension Reduction

Prior to building and fitting the Linear Regression (LR) model, Dimension Reduction (DR) will be performed on the dataset to ensure that there is not collinearity or multicollinearity present, as per LR models' assumptions. Before conducting the DR, a number of checks will be done to ensure that DR can be performed on the dataset.

Correlation matrix

A correlation matrix was constructed and inspected for the variables in the dataset to identify clusters and relationships between variables. A graph was also generated to present the relationships between all values and the significance of these relationships. This is done by including the coefficient of the variable relationship, and marking the matrix cell with an "X" if it is an insignificant relationship.



Several clusters were found in the graph, with a number of coefficients having strong relationships. A cluster was found that containing almost singular relationships, with coefficients over 0.9. These strong relationships across the correlation matrix suggests the presence of ***multicollinearity***. The variables causing this multicollinearity are *Percentile*, *Quartile*, and *Second Decile*, which are all derived variables. These variables will be removed from the dataset as their coefficients (> 0.9) are too high to be used for DR. We also remove our y variable, which is ***G_SC***. Since the objective of the DR is reduce the number of variables, the ***Principle Component Analysis (PCA)*** was chosen over the ***Factor Analysis (FA)***.

Probability and Statistical Inference
Continuous Assessment Part II

Data Suitability Check

Following the inspection of the correlation matrix and removing the variables with coefficients that are too high, the data must be checked to see if it is suitable to be used in DR. This is done using a few measures.

KMO

The first is Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO), which measures how well suited the data is for DR. KMO must be greater than 0.5 for the data to be considered suitable, but the higher it is, the more suitable DR will be for the data. The general descriptors set by Kaiser for the KMO are: 0.00 to 0.49 is unacceptable, 0.50 to 0.59 is miserable, 0.60 to 0.69 is mediocre, 0.70 to 0.79 is middling, 0.80 to 0.89 is meritorious, and finally 0.90 to 1.00 is marvellous. The KMO of the dataset was found to be 0.88, which is considered to meritorious according to Kaiser.

Bartlett's Test of Spherity

The second measure is Bartlett's Test of Spherity, which checks to see if the data can be combine together. This must have p-value > 0.05. The Bartlett's Test resulted in p-value > 0.05; therefore the data can be combined together.

The Determinant

The third measure is the Determinant, which measures the multicollinearity in the data. If the Determinant is greater than 0.00001, the data can be used. The determinant result was 0.0014, which is greater than the cut-off value, meaning that the data can be used.

Communalities and Loadings

The resultant communalities were uniformly high (> 0.8), meaning that the variance of the variables was accounted for in the components. The loadings, which are the correlation between the factor and the manifest variables are mostly < 0.3.

Section 3.2 – Model

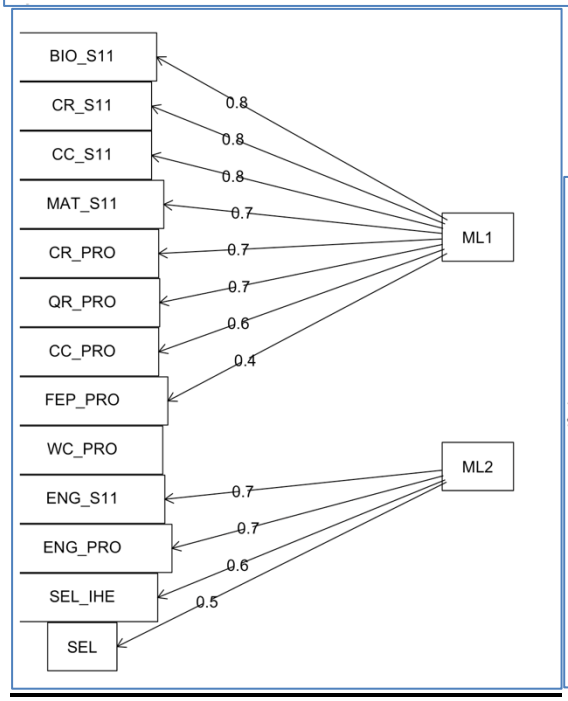
Section 3.2.1 Statistical Evidence

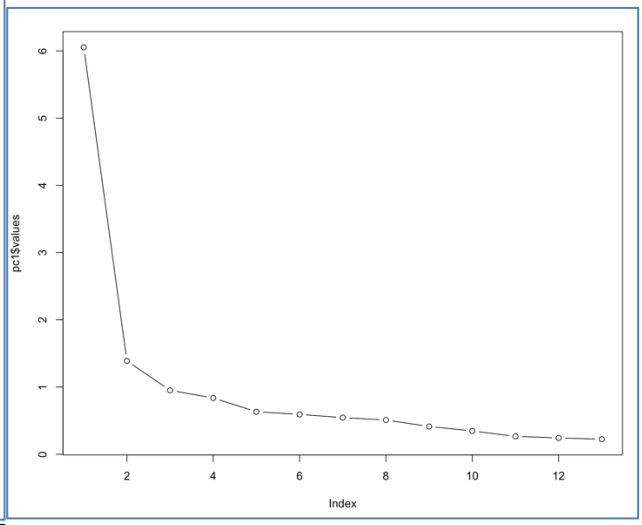
A principal component analysis (PCA) was conducted on the 13 numerical variables with orthogonal rotation (varimax). Bartlett's test of sphericity, $X2(78) = 81233.41$, $p < .001$, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Two components had eigenvalues over Kaiser's criterion of 1 and in combination explained 65.02% of the variance. The scree plot was slightly ambiguous and showed inflexions that would justify retaining either 2 or 3 factors. Given the large sample size, and the convergence of the scree plot and Kaiser's criterion on three components, two components were retained in the final analysis. The 1st Component represents the student's grades, and the 2nd component represents the student's social class. The student's grades all had high reliabilities, all Cronbach's $\alpha = .76$. However, the student's social class component had a relatively low reliability, Cronbach's $\alpha = .55$.

Probability and Statistical Inference
Continuous Assessment Part II

```
Reliability analysis
Call: alpha(x = socialClass)

  raw_alpha std.alpha G6(smc) average_r S/N   ase mean  sd median_r
      0.55      0.78    0.76      0.47 3.5 0.002   34 9.6     0.42

 lower alpha upper     95% confidence boundaries
0.55 0.55 0.56

 Reliability if an item is dropped:
       raw_alpha std.alpha G6(smc) average_r S/N alpha se   var.r med.r
ENG_S11     0.089      0.67    0.57      0.40 2.0   0.0014 0.00138  0.42
ENG_PRO     0.165      0.69    0.60      0.42 2.2   0.0023 0.00098  0.42
SEL         0.600      0.78    0.74      0.54 3.5   0.0023 0.03002  0.46
SEL_IHE     0.600      0.75    0.71      0.50 3.0   0.0023 0.04409  0.39

 Item statistics
            n raw.r std.r r.cor r.drop mean    sd
ENG_S11 12411  0.89  0.84  0.81   0.75 61.8 14.30
ENG_PRO 12411  0.96  0.81  0.77   0.74 67.5 25.50
SEL     12411  0.43  0.70  0.52   0.40  2.6  1.11
SEL_IHE 12411  0.49  0.74  0.59   0.47  2.4  0.93

Non missing response frequency for each item
            1    2    3    4 miss
SEL     0.17 0.38 0.12 0.33    0
SEL_IHE 0.09 0.62 0.07 0.22    0
```

```
         eigenvalue variance.percent cumulative.variance.percent
Dim.1  2732.8871119     45.894431305                    45.89443
Dim.2  1139.2788593     19.132350954                    65.02678
Dim.3   772.4380039     12.971850449                    77.99863
Dim.4   425.2980847      7.142195389                    85.14083
Dim.5   326.9453057      5.490519095                    90.63135
Dim.6   255.6331894      4.292947117                    94.92429
Dim.7   127.8133311      2.146418751                    97.07071
Dim.8    62.4291076      1.048396173                    98.11911
Dim.9    53.6814047      0.901492612                    99.02060
Dim.10   29.6882429      0.498566157                    99.51917
Dim.11   26.9786537      0.453062977                    99.97223
Dim.12    1.1228565      0.018856564                    99.99109
Dim.13    0.5307122      0.008912455                   100.00000
```





Hypotheses

1. ***Not all 15 variables affect the target variable***

This hypothesis was found to be true. The PCA resulted in 2 components which are a linear combination of 11 variables, not 15.

2. ***Citizen Competencies will not have an effect on the Global Score***

This hypothesis was found to be false. A significant relationship was found between both variables using a Pearson correlation test, since both variables are continuous and normal ($p < 0.05$, $df = 12409$, coefficient = 0.757).

3. ***The English Score in Saber Pro (University) is highly dependent on the English Score in Saber 11 (Grade 11)***

This hypothesis was found to be true. A significant relationship was found between both variables. A parametric Pearson correlation test was used as both variables are continuous and follow a normal distribution ($p < 0.05$, $df = 12409$, coefficient = 0.738).

Probability and Statistical Inference
Continuous Assessment Part II

4. ***A student's social class affects his/her ability to speak English positively***

This hypothesis was found to be true. A significant relationship was found between both variables. Because Social class is a nominal variable, a Spearman correlation test was used ($p < 0.05$, rho = 0.3899755).

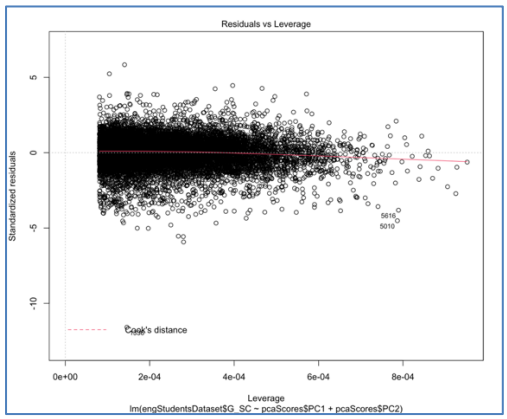
These variables will now be used in the Linear Regression model.

Section 3.2.2 Model

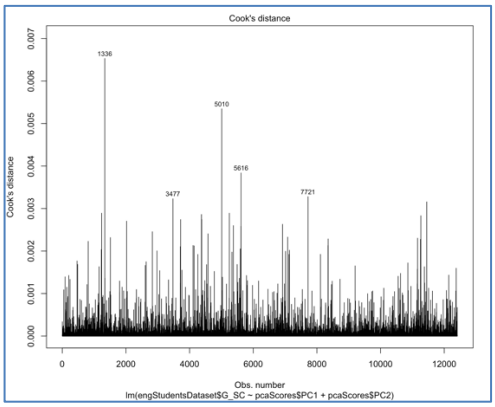
Linear Regression models use continuous variables that have linear relationships. The assumptions of a linear regression model are: (1) absence of outliers, (2) constant, non-zero variation, (3) homoscedasticity, (4) normal distribution of errors, (5) absence of collinearity and multicollinearity, and (6) linearity.

Outliers

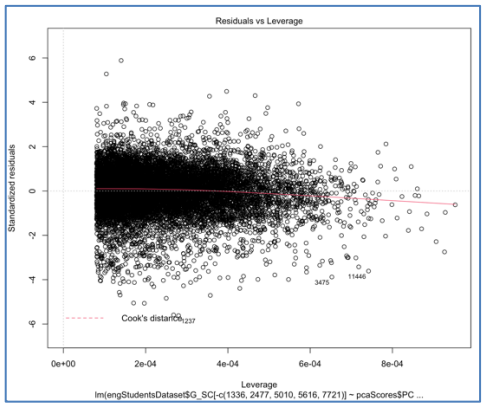
Linear Regression assumes is sensitive to outliers as it can alter the model's result. Accordingly, it assumes that the model has no outliers. The Residuals vs Leverage plot is used to show outliers or extreme values. The graph showed many outliers present with the observation 1336 that is outside of the range of other values. Cook's Distance was graphed to find the 5 most extreme values in the model and these were removed. This fixed the outlier issue.



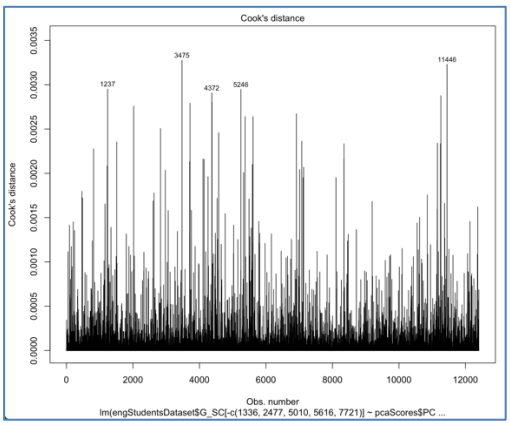
Model 1



Model 1



Model 2

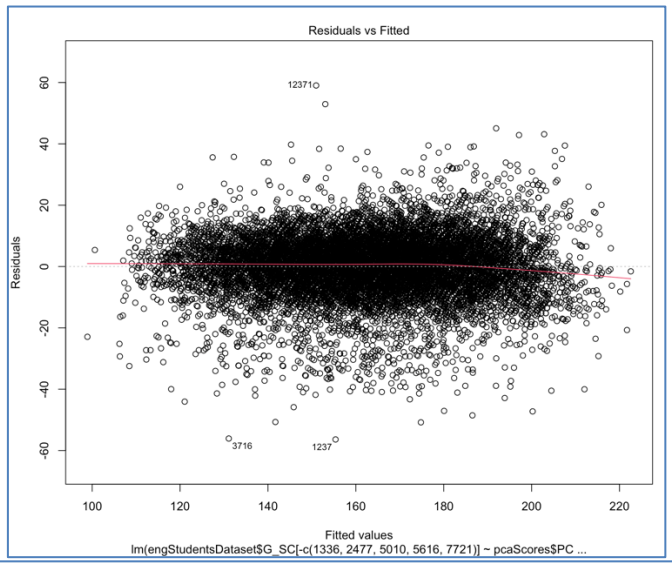


Model 2

Probability and Statistical Inference
Continuous Assessment Part II

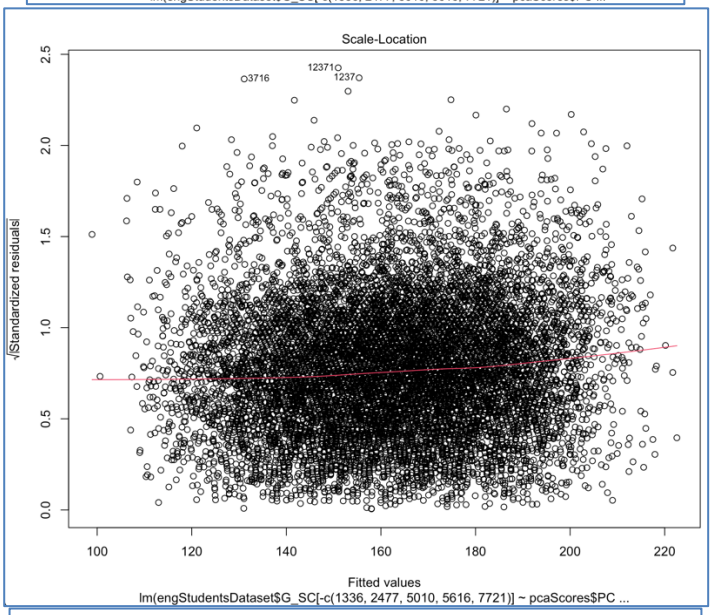
Linearity

Linear Regression assumes linear relationships between the predictors. This means that the Residuals vs Leverage graphs should show a random pattern as is the case in the graph on the right. This means the linearity assumption holds.



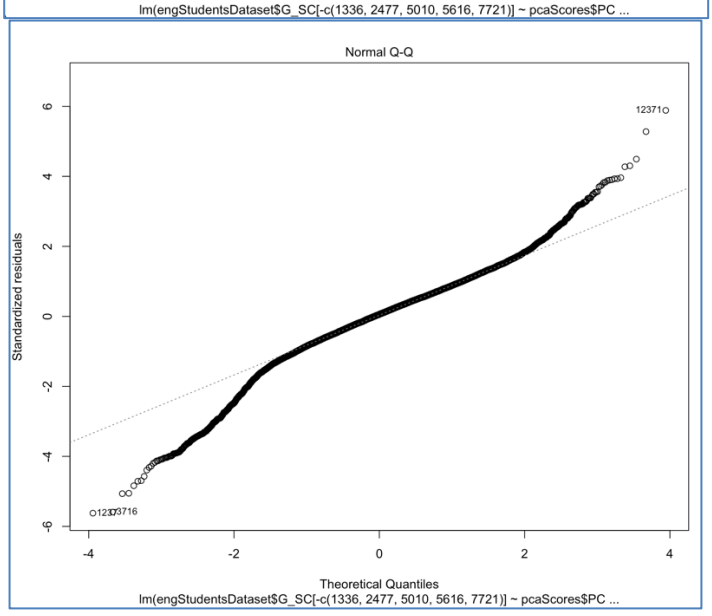
Homogeneity of Variance

Linear Regression assumes that the variance of is homogenous. The graph on the right checks to see that the values are spread along the predictors' ranges. The graph shows an almost horizontal line, which means we can assume homoscedasticity.



Normality of Residuals

Linear Regression assumes that the residuals follow a normal distribution. This means that when plotted on a QQ Plot, the should almost all be on the same line. This is not the case in our model as the plot shows that the residuals are not normally distributed. The assumption of normality does not hold.



Probability and Statistical Inference
Continuous Assessment Part II

Collinearity

Collinearity was addressed using Principal Component Analysis (PCA), which showed that the 17 numeric columns can be reduced to 2 uncorrelated components.

Reporting in APA

A multiple regression analysis was conducted to determine how much the Global Score is affected by the Principal Components, which are a linear combination of MAT_S11, CR_S11, BIO_S11, ENG_S11, QR_PRO, CR_PRO, CC_PRO, ENG_PRO, FEP_PRO, SEL, SEL_IHE.

Examination of the histogram, normal P-P plot of standardised residuals and the scatterplot of the dependent variable, Global Scores, and standardised residuals showed that the some outliers existed. After plotting the Cook's Distance the 5 most extreme observations, were found to have Cook's Distance > 0.003. These values were removed. They were found to have in impact on the model (Model1, Adjusted R2 = 80.8%, F-statistic with $p < 0.05$, statistically significant model) (Model2, Adjusted R2 = 81.1%, F-statistic with $p < 0.05$, statistically significant model).

Examination for multicollinearity showed that the tolerance and variance influence factor measures were within acceptable levels (VIF <2.5). The scatterplot of standardised residuals showed that the data met the assumptions of homogeneity of variance and linearity. The data however did not meet the assumption of normally distributed residuals.

Section 4 – Discussion/Conclusion

This report set out to answer the research question:

How far can we reduce the dimensionality of the dataset while retaining the variability in the target variable (Global Score)?

A Principal component analysis was conducted and revealed that the 17 available numeric variables can be reduced to 2 component, Grades and Social Skills. These components contain 11 of the original variables. The following hypothesis were defined prior to the beginning of the analysis and then were tested.

1. ***Not all 15 variables affect the target variable:*** true.
2. ***Citizen Competencies will not have an effect on the Global Score.*** false.
3. ***The English Score in Saber Pro (University) is highly dependent on the English Score in Saber 11 (Grade 11).*** true
4. ***A student's social class affects his/her ability to speak English positively***. true

Linear Regression was then performed to reveal the effect of the components on the target variable, G_SC. The assumptions were first checked, and they were found to all hold true, except for the assumption of normally distributed residuals. A significant model was produced with an adjusted R2 value of 81.1%.

While this research showed reliable results, it contains limitations. The residuals in the Linear Regression model did not follow a normal distribution, which may affect the results of the model.